UNIVERSIDADE FEDERAL DE SANTA CATARINA

CENTRO DE COMUNICAÇÃO E EXPRESSÃO

CURSO DE PÓS-GRADUAÇÃO EM LINGÜÍSTICA

LETÍCIA LUISE KRIEGER STEIN

O USO DO VERBO *FICAR* EM LÍNGUA PORTUGUESA: UMA PESQUISA BASEADA EM *CORPUS*

Dissertação apresentada ao Programa de Pós-Graduação em Lingüística da Universidade Federal de Santa Catarina, como requisito parcial para a obtenção do título de Mestre em Lingüística

Orientador: Prof. Dr. Marco Antonio Esteves da Rocha

Florianópolis, junho de 2004.

O USO DO VERBO *FICAR* EM LÍNGUA PORTUGUESA:UMAPESQUISA BASEADA EM CORPUS

LETÍCIA LUISE KRIEGER STEIN

BANCA EXAMINADORA:

_____

Prof. Dr. Marco Antonio Esteves da Rocha

Orientador

_____

Prof. Dr. A. P. B. Sardinha

_____

Profa. Dra. Edair Görski

_____

Profa. Dra. Luizete Guimarães Barros

(suplente)

AGRADECIMENTOS

Ao CNPQ , órgão financiador dessa pesquisa.

Aos professores e colegas do curso de Pós-Graduação em Lingüística.

Ao Prof. Dr. Marco Antonio Rocha, meu orientador, pela disponibilidade e orientação que tornou possível a realização desse projeto .

À minha família.

SUMÁRIO

RESUMO

ABSTRACT

CAPÍTULO 1 - INTRODUÇÃO

CAPÍTULO 2 - REFERENCIAL TEÓRICO

2.1. LINGÜÍSTICA DE CORPUS

2.1.1 Definição de *corpus*

2.1.2. Um breve histórico da Lingüística de Corpus

2.1.3. Os corpora disponíveis no mercado

2.1.4. Objetivos e métodos da pesquisa com base em *corpus*

2.1.5. A ligação entre a Lingüística Computacional e a Lingüística de Corpus

2.1.6. Metodologia de corpus e o uso de corpora para a pesquisa lexicográfica

CAPÍTULO 3 - METODOLOGIA

3.1. CONTEXTUALIZANDO A PESQUISA

3.1.1 Descrevendo o cenário da pesquisa

3.1.2 Objetivos da pesquisa e hipóteses de trabalho

3.1.3. Definindo a polissemia

3.1.4 A lingüística cognitiva

CAPÍTULO 4 – A PESQUISA

4.1 A coleta de dados

4.2. Resultados: descrição e análise

CONCLUSÃO

BIBLIOGRAFIA

ANEXOS

RESUMO

Esta pesquisa objetivou descrever uma análise a partir de corpus do uso do verbo **ficar**, a fim de comparar o seu uso na língua portuguesa e aqueles propostos por um dicionário dessa, o *Dicionário Houaiss da língua portuguesa* (2001, 1 ed.- doravante citado como: Dicionário Houaiss), e verificar se o dicionário dá conta da polissemia do verbo. Os dados foram coletados no corpus organizado pelo NILC (Núcleo Interinstitucional de Lingüística Computacional), o qual é constituído por uma variedade de gêneros textuais suficiente para os propósitos da pesquisa. Foram encontradas 43529 ocorrências do verbo ficar, em 42 formas diferentes. Do total de ocorrências do verbo no corpus foi extraída uma amostra com 500 delas, as quais foram analisadas com suas variações lexicais e semânticas, com o objetivo final de comparar os resultados encontrados com aqueles propostos pelo dicionário selecionado para a pesquisa.

ABSTRACT

This study aims at describing a corpus based analysis of the verb *ficar* (to stay) comparing its usage to that one proposed by the Houaiss dictionary of Portuguese language to verify if this dictionary suffices all the usages of the selected verb. The data was taken from the corpus collected by NILC (Núcleo Interinstitucional de Lingüística Computacional) which contains enough variety of texts for the purposes of this study. In the corpus were found 43529 occurences of the selected verb in 42 different usages. From all the occurences that were found in the corpus it was extracted a sample with 500 occurences of the verb. The sample was analysed under the lights of the lexical and semantic variations of each occurrence. The objective of this analysis was to compare the results with what is proposed by the Houaiss Dictionary of Portuguese language.

CAPÍTULO 1 – INTRODUÇÃO

Este trabalho de pesquisa objetivou investigar o uso do verbo *ficar* em língua portuguesa, utilizando as metodologias de análise da Lingüística de Corpus e os recursos para a manipulação de corpus da Lingüística Computacional. A abordagem com base em corpus representa uma opção científica e metodológica e considera o corpus como base a partir da qual se desenvolve toda a análise da língua; os programas de manipulação de corpus são uma conseqüência disso, assim como os corpora digitalizados são uma conseqüência do surgimento do computador. O corpus computadorizado fez-se então necessário, para que pudesse ser manipulado, utilizando-se as ferramentas do programa de manipulação de corpus *WordSmith*.

A escolha do verbo *ficar* deu-se haja vista seus diversos usos encontrados na língua portuguesa, surgindo o interesse de analisar seu emprego – como e com que freqüência ele é mais utilizado –, a fim de se comparar seu emprego na língua portuguesa e as propostas por um dicionário, que o normatiza. Para essa atividade, foi tomado por base o Dicionário Houaiss, dentre os mais respeitados do país. Como foi mencionado anteriormente, fez-se necessário o uso de um corpus para o desenvolvimento da pesquisa, com textos em língua escrita. Optou-se, então, pelo corpus organizado pelo NILC

(Núcleo Interinstitucional de Lingüística Computacional), que é um corpus acessível, constituído por uma variedade de gêneros textuais suficiente para os propósitos do trabalho, e dada também sua representatividade.

O corpus digitalizado permite o uso de recursos de busca e recuperação intrínsecos ao processamento de dados em computador, e constitui uma abordagem distinta dentro da lexicografia, a qual ficou conhecida a partir da publicação dos dicionários e de outros trabalhos da série COBUILD. Na década de 20, Thorndike fez um levantamento e um estudo das palavras mais freqüentes da língua inglesa. Seu estudo influenciou trabalhos posteriores, dentre os quais o *General Service List of Enlglish Words*. Hoje em dia, com o advento do computador, e com os programas de manipulação de corpus, os pesquisadores manipulam corpora com milhões de palavras. Os corpora eletrônicos disponíveis no mercado atualmente são em grande número; alguns deles como o Brown Corpus (Brown University Standard Corpus of Present-day American English), constituem marcos de referência do uso de corpora para a pesquisa lingüística.

A população do verbo *ficar* no corpus de estudo é muito grande, 43525 ocorrências em 42 formas diferentes. Para que todas elas fossem analisadas e comparadas, seria necessária uma grande demanda de tempo, o que não era o caso, em se tratando de uma dissertação de mestrado. Por isso optou-se pelas amostras, a fim de se possibilitar um levantamento estatístico e análise dos dados encontrados. O fator ‘tamanho’, no entanto, é algo muito importante na Lingüística de Corpus; em outras palavras, quanto maior for o corpus mais representativo ele é. Quanto maior, mais chance

há de que formas raras da língua, aquelas que apresentam freqüência 1, sejam 'escolhidas' na compilação do corpus. No Capítulo 3, o leitor terá acesso a uma tabela onde se encontram listadas as formas do *ficar* no corpus, e algumas delas aparecem apenas 1 vez, em meio a milhões de palavras.

A pesquisa encontra-se dividida em duas partes: a teórica e a prática. A primeira, Capítulos 2 e 3, apresenta o resultado do estudo da literatura que definiu os rumos da pesquisa. O Capítulo 2 traz os conceitos sobre a Lingüística de Corpus, seu histórico, objetivos e métodos. Ao longo dos anos, muitas pesquisas já foram desenvolvidas utilizando a Lingüística de Corpus e, conseqüentemente, a abordagem com base em corpus, conforme já exposto brevemente. Tal opção mostra uma preocupação por parte dos lingüistas e uma tendência em se estudar a língua em uso, ao invés de utilizar a abordagem com base em regras, cujos resultados são fundamentados no estudo de sistemas de regras/ estruturas artificialmente estabelecidos, não se pretendendo descartar tal abordagem do cenário da pesquisa lingüística.

A aplicação da abordagem com base em corpus para o estudo do verbo *ficar* encontra-se detalhada no Capítulo 3. A escolha do corpus envolveu a análise de alguns princípios discutidos pelo lingüista de corpus Beber Sardinha (2000): a autenticidade do corpus e sua representatividade. Autenticidade, no sentido de os textos terem sido produzidos para outros fins que não a pesquisa lingüística, e tais textos devam ter sido escritos por falantes nativos. A representatividade diz respeito ao tamanho do corpus; ele

deve ser grande o suficiente para conter o maior número possível das formas (verbais, nominais, etc.) da língua em questão.

A busca específica pelo verbo *ficar* no corpus foi feita em três etapas, utilizando-se o programa de manipulação de corpus *WordSmith*. (Para mais detalhes sobre o programa, ver seção 2.1.6) Na primeira busca objetivou-se fazer um levantamento estatístico sobre a presença do verbo no corpus. Em seguida, procedeu-se a uma busca, respeitando-se o limite do *WordSmith*, que é de apresentar 16000 ocorrências do termo ou expressão desejados. Dessa população extraiu-se uma amostra com 500 ocorrências do verbo, na qual se baseiam a análise e os resultados da pesquisa. Os resultados encontram-se tabelados.

A última etapa da pesquisa, detalhada no Capítulo 4, envolveu o estudo do verbo: a análise das variações lexicais – um levantamento de como é marcada a presença do verbo no corpus – e a análise semântica – os sentidos que o verbo assume –, buscando-se estabelecer um padrão entre as duas variações: se o complemento do verbo exerce alguma influência sobre o seu sentido. Outro elemento analisado foi o contexto em que o verbo apareceu, também buscando-se estabelecer alguma relação com a sua polissemia. Finalmente, os resultados encontrados foram comparados ao que é normatizado pelo Dicionário Houaiss.

Finalmente, na Conclusão encontra-se uma reflexão sobre os resultados da pesquisa, tanto sobre a escolha metodológica quanto sobre a análise do corpus, seguida pela indicação bibliográfica.

CAPÍTULO 2 – REFERENCIAL TEÓRICO

2.1 LINGÜÍSTICA DE CORPUS

2.1.1 Definição de corpus

O primeiro registro conhecido do uso da palavra corpus no sentido de coletânea ou corpo de textos, data do século VI, quando o imperador Justiniano incentivou a compilação de uma coletânea de leis e princípios legais romanos chamada *Corpus Juris Civilis*. Duzentos anos depois, no século VIII, foi compilada uma listagem de palavras latinas difíceis, em ordem alfabética, o *Corpus Glossary*, cuja finalidade se aproxima mais da de um corpus lingüístico do que o primeiro. (Para uma discussão mais aprofundada, ver Rocha, 2000).

Em termos gerais, um corpus pode ser definido como uma coletânea de textos reunidos para diversos fins de pesquisa lingüística. Tendo sido concluído, são vastas as suas opções de uso, indo além das previstas na sua fase de compilação. No âmbito da lingüística de corpus, um corpus é “uma coletânea de textos, que se pressupõe seja representativa de uma língua, dialeto, ou outro subconjunto dessa língua, a ser utilizada

para fins de análise lingüística."(Francis, 1992). O crescente interesse pelo uso do corpus para a pesquisa lingüística está relacionado com o desenvolvimento do corpus computadorizado, que é manipulado com a ajuda de programas de computador. Segundo Stig Johansson (1998:3):

"The place of corpora has been controversial in liguistics. Yet there is now more interest than ever in the use of corpora in language research. This is connected with the increasing interest among linguists in language use rather than language systems in the abstract. It is no doubt also due to the development of corpora in machine-readable form and of techniques for exploring and analysing such corpora."

Johansson também afirma que os usos que se pode fazer de um corpus dependem das escolhas metodológicas realizadas pelo pesquisador quando no processo de seleção e preparação do material utilizado na compilação do corpus. O uso de corpora e as abordagens utilizadas para a pesquisa lingüística, o uso de um corpus computadorizado e de programas de manipulação de corpus serão assuntos discutidos nas seções a seguir.

2.1.2 Um breve histórico da lingüística de corpus

A lingüística de corpus é dividida em dois momentos: a.c. (antes do computador) e a de nossos dias, já relacionada à lingüística computacional. Starnes e Noyes (1946) discutem, em seu estudo da história da lexicografia da língua inglesa, a prática de se coletarem ocorrências de palavras em textos diversos para a compilação de dicionários.

Samuel Johnson também discute esse método, em *Plan of an English Dictionary* (1747) e no prefácio à edição de 1755 do *Dictionary of the English Language*. O autor selecionava as ocorrências das palavras (em textos/artigos publicados) para as quais seriam elaborados os verbetes. Seus amanuenses copiavam as citações em pedaços de papéis, para posterior análise pelo pesquisador. A técnica demandava muito trabalho, tanto da parte do pesquisador quanto da dos amanuenses, e Abercrombie (1965) definiu o método de pesquisa com base em corpus da sua época e da de seus antecessores como um "pseudoprocedimento", pois "embora não seja talvez literalmente impossível,... seria uma maneira tão árdua e trabalhosa de conduzir uma investigação, que ninguém em seu juízo perfeito se disporia a utilizá-la".

No segundo momento, o advento do computador e a sua utilização sistemática em diferentes áreas de pesquisa lingüística fez com que a Lingüística de Corpus ficasse inevitavelmente relacionada à Lingüística Computacional e ganhasse dimensões maiores, como a manipulação de corpora de textos de grande porte. Atualmente, a busca por um item em um corpus de milhões de palavras faz-se muito rapidamente, e os resultados são bastante confiáveis, graças à capacidade dos computadores de buscar, recuperar, selecionar e calcular conteúdos. Sendo assim, o computador, e o seu decorrente uso para pesquisa lingüística, pôs fim a limitações enfrentadas pelos lingüistas que se dispunham a utilizar a abordagem com base em corpus, como investigar corpora com um grande número de palavras.

Dentre os possíveis usos de um corpus para a investigação lingüística, Johansson (1998) destaca a importância do uso de corpora para o surgimento de novas descrições lingüísticas e propõe um modelo integrado do dicionário, da gramática e do corpus. Em tal modelo os dicionários e gramáticas seriam um guia da língua em uso, ao invés de uma coleção de exemplos fora de contexto:

"Traditionally, lexicographers and grammarians were severely restricted by considerations of space. What we can have now is an open-ended dicitonary which is linked to a corpus. Similarly, a grammar can be linked to a corpus. The dictionary and the grammar provide the description, and the corpus gives examples of language use in context. (…) Corpora may thus serve to unite fields of study that have traditionally been kept apart, and the result will be a coherent language description."

O modelo proposto por Johansson mostra uma tendência em se utilizar a língua em uso para a descrição lingüística, ao invés de se empregarem subconjuntos artificialmente delimitados das linguagens naturais influenciados por sistemas de regras/estruturas estabelecidas pela gramática normativa. Enquanto um corpus permite ao pesquisador analisar a língua em uso, com resultados que se aproximam da realidade dos falantes/usuários da língua, os sistemas de regras/ estruturas processam frases a fim de separar as gramaticais das não gramaticais, o que se encaixa nas estruturas pré-estabelecidas e o que, conseqüentemente, não se encaixa, criando assim modelos artificiais da língua em estudo.

O estudo da língua utilizando os sistemas de regras/ estruturas recebe, dentro da lingüística computacional, o nome de abordagem com base em regras. Os seguidores dessa abordagem acreditam que, se o objetivo é aproximar a capacidade lingüística da máquina à de um ser humano, por exemplo, a melhor maneira de consegui-lo é por meio da reprodução do funcionamento da língua humana. Sabe-se, porém, que o processamento de linguagens naturais (PLN) está além da capacidade das máquinas. (Para mais detalhes, ver Rocha, 2000)

São duas, então, as abordagens no cenário da investigação lingüística, no caso da lingüística computacional: a abordagem com base em regras (que faz uso de regras gramaticais e de outros conceitos para a investigação lingüística) e o corpus. A abordagem com base em regras, conforme exposto acima, tende a privilegiar subconjuntos artificialmente delimitados das linguagens naturais. Ela sofre influência das gramáticas gerativas, que reverenciam somente as construções bem definidas da língua. "Nas diversas tentativas de implementação feitas com base nestas gramáticas, porém, fica caracterizada a enorme distância que ainda existe entre estes subconjuntos de uma linguagem natural e a diversidade das línguas conforme empregadas para fins de comunicação na vida real." (Rocha, 2000).

Enquanto a abordagem com base em regras analisa 'subconjuntos artificialmente delimitados', a abordagem com base em corpus permite ao pesquisador analisar um grande número de ocorrências de palavras a partir de textos que as pessoas lêem e escrevem diariamente. O *Oxford English Dictionary*, publicado pela primeira vez em

1884, e o *Webster´s New International Dictionary, Second Edition*, publicado pela Merriam-Webster Co., em 1934, registram o uso de citações de papel na sua elaboração: os dados eram anotados em pedaços de papel e, posteriormente, analisados para a elaboração do material em questão.

Muitos dos trabalhos produzidos atualmente manipulam corpora com milhões de palavras e fazem uso de modelos estatísticos complexos para chegar a seus resultados. O uso do corpus computadorizado possibilita ao pesquisador alcançar resultados bastante confiáveis e, com o uso de programas de manipulação de corpus, ele pode incorporar à análise lingüística as noções de freqüência e probabilidade. Os computadores permitem o armazenamento e manipulação de corpora maiores, como é o caso do *Bank of English* e do NILC, com textos em língua escrita.

Por prover recursos metodológicos e filosofias diferentes para a descrição da língua, a lingüística de corpus é uma metodologia que faz parte de uma abordagem de pesquisa, uma 'filosofia' de investigação lingüística, e não um ramo específico da lingüística, relacionando-se às mais diversas áreas. Até o surgimento do computador, muito da análise lingüística era fundamentada na intuição do lingüista. O corpus serve para testar e comprovar, ou não, o que é intuído sobre a língua. Segundo Leech (1991), a pesquisa com base em corpus deve ser vista como corpus **e** intuição trabalhando juntos, ao invés de corpus **ou** intuição. Subentende-se que, para ter intuição ou conhecimento da língua a ser pesquisada, o pesquisador é falante nativo; além disso, precisa ter

conhecimento **sobre** a língua, no caso, ser um lingüista. Em termos gerais, podemos descrever a pesquisa com base em corpus desta forma:

"A researcher has an intuition about language, checks this against the data the corpus provides, and this checking process frequently suggests other avenues of research to be taken, often entirely unsuspected at the start of the process (the socalled "serendipity" principle, Higgins, 1991). (Partington, 1998)"

Assim, como temos alguns elementos trabalhando juntos, intuição/conhecimento lingüístico/corpus, a análise lingüística que utiliza o corpus computadorizado nasce também de uma busca por padrões recorrentes, freqüência ('passado') e probabilidade ('futuro'): se algo ocorre freqüentemente na língua, é significativo. A freqüência serve como base para a probabilidade, para prever como a língua pode vir a se comportar e, finalmente, como o discurso é construído globalmente. Alguns dados que às vezes passam despercebidos à intuição do pesquisador tornam-se explícitos com o uso de concordanciadores, programas de gerenciamento de corpus que apresentam listas de freqüência e ocorrência de palavras ou de uma palavra específica dentro de um corpus.

A disponibilidade no mercado, como visto a seguir, e a facilidade do uso de um corpus fez com que seu uso abrangesse várias áreas de pesquisas lingüísticas, dentre as quais estudos de estilo e autoria, estudos históricos (Lingüística Diacrônica), estudos de tradução, de registros, léxico e sintaxe. Tal abrangência decorre do fato de que a

Lingüística de Corpus é uma abordagem metodológica de análise e investigação de fenômenos lingüísticos, com os seguintes traços característicos (Leech, 1992):

- foco no desempenho lingüístico, ao invés da competência;
- foco na descrição lingüística, ao invés dos universais lingüísticos;
- foco tanto na quantidade quanto nos modelos qualitativos de linguagem;
- foco na pesquisa empírica, ao invés de uma visão mais racionalista de pesquisa científica.

2.1.3 Os corpora disponíveis no mercado

Atualmente são muitos os corpora de língua escrita ou falada disponíveis no mercado, que podem ser utilizados para os mais diversos fins de pesquisa lingüística, como discutido anteriormente. Os corpora de língua falada, entretanto, tendem a ser menores do que os corpora de língua escrita. Apesar disso, o *Cancode corpus*, que está sendo desenvolvido pela *Nottingham University* e pela *Cambridge University Press*, contém aproximadamente 10 milhões de palavras de 'falas' transcritas. Dos corpora comerciais de língua escrita, o que chama a atenção é o *BNC (British National Corpus)*, do seu total de palavras, 10% representam uma população de língua falada, produzido por uma parceria entre a *Oxford University Press*, *Longman*, *Chambers Harrap*, as universidades de *Oxford* e *Lancaster* e *British Library*. O referido corpus conta com aproximadamente 100 milhões de palavras.

Os centros de pesquisa em Lingüística de Corpus são, atualmente, em grande número, sendo que os mais renomados encontram-se na Grã-Bretanha (Birmingham, Lancaster, Liverpool, etc). A partir da tabela abaixo (1) podem-se observar os corpora compilados ou em compilação, e o crescimento do número de corpora na no meados de 1990 (Sardinha, 2000: 330). Como a tabela data refere-se à década de 90, alguns dos dados não estão atualizados.

Tabela 1- *Os corpora disponíveis no mercado*

Corpus	Lançamento/ Referência na Literatura	Palavras	Composição
Brown Corpus (Brown University Standard Corpus of Present-day American English)	1964	1 milhão	Inglês americano escrito
AHI (American Heritage Intermediate Corpus)	1971	5 milhões	Inglês americano escrito
LOB (Lancaster-Oslo-Bergen)	1978	1 milhão	Inglês britânico escrito
LLC (London-Lund Corpus)	1980	500 mil	Inglês britânico falado
Birmingham Corpus (Birmingham University International Language Database)	1987	20 milhões	Inglês britânico

Kolhapur Corpus (Indian English)	1988	1 milhão	Inglês indiano escrito
TOSCA Corpus (Tools for Syntactic Corpus Analysis)	1988	1,5 milhão	Inglês britânico escrito
SEU Corpus (Survey of English Usage)	1989	1 milhão	Inglês britânico, escrito e falado
CHILDES (Child Language Data Exchange)	1990	20 milhões	Inglês infantil falado
Nijmegen Corpus	1991	132 mil	Inglês britânico, escrito e falado
Map Task Corpus	1991	147 mil	Inglês escocês falado
LCLE (Longman Corpus of Learner's English)	1992	10 milhões	Inglês escrito por estrangeiros
SEC (Lancaster/IBM Spoken English Corpus)	1992	53 mil	Inglês britânico falado
Wellington Corpus (Written New Zealand English)	1993	1 milhão	Inglês neozelandês escrito
POW (Polytechny of Wales Corpus)	1993	65 mil	Inglês infantil falado
Wellington Corpus of Spoken New Zealand English	1995	1 milhão	Inglês neozelandês falado
BNC (British National Corpus)	1995	100	Inglês britânico,

		milhões	escrito e falado
ICLE (International Corpus of Learner English)	1997	200 mil**	Inglês escrito por estrangeiros
Bank of English	1997	320 milhões	Inglês britânico

* Previsão

** Cada variedade nacional

Sardinha (2000: 330) destaca os corpora Brown, Birmingham e BNC como "marcos de referência histórica". E continua: "o corpus Brown é um marco por razões óbvias: é pioneiro. O corpus Birmingham é importante porque foi o primeiro a ultrapassar a marca de 1 milhão de palavras iniciada pelo Brown. Vale lembrar que o corpus Birmingham se tornaria o Bank of English, sempre em crescimento, atingindo agora 320 milhões de palavras. Por fim, o BNC é um marco histórico porque foi o primeiro a conter 100 milhões de palavras, e ainda é, dentre os mega-corpora, o único disponível para compra (dentro da Comunidade Européia apenas)[1]. O Bank of English é de acesso restrito aos pesquisadores ligados ao COBUILD e à editora Collins."

Os corpora em língua inglesa, disponíveis no mercado, são em grande número. No Brasil, os dados sobre a lingüística de corpus mostram que ela se encontra em desenvolvimento, com os estudos dirigidos para as áreas de PLN (Processamento de

[1] O BNC já se encontra disponível para compra fora da Comunidade Européia.

Linguagens Naturais), Lexicografia e Lingüística Computacional. (Para mais detalhes, ver Sardinha, 2000).

Na língua portuguesa já existem cinco corpora de referência nacional:

NILC (trabalho conjunto da USP- São Carlos e UFSCar);

CRPC (corpus de referência do português contemporâneo);

PORTEXT (Banco de português);

Tycho- Brahe (português histórico);

Corpus Natura.

VARSUL (projeto interisntitucional que integra a UFPR, UFRGS, a PUC-RS e a UFSC)

Banco de Português (o maior corpus de língua portuguesa.)

Grupos empresariais internacionais têm mostrado interesse no estudo baseado em corpora, o que gerou parcerias entre eles e universidades. A série COBUILD (Birmingham University e editora Collins) é um exemplo da parceria entre empresas e universidades. Outras empresas, como a Microsoft, Xerox e Cânon, financiam pesquisas com finalidades comerciais, como a informatização de reconhecimento de grandes bases de dados e a montagem de sistemas inteligentes de reconhecimento de voz e gerenciamento de informação. O surgimento do computador e a disponibilidade de corpora em meio eletrônico estão ligados ao desenvolvimento da Lingüística de Corpus.

2.1.4 Objetivos e métodos da pesquisa com base em corpus

Os estudos da língua podem enfatizar tanto sua estrutura quanto seu uso.(Para mais detalhes, ver o item 2.1.2) Segundo Biber, Conrad et al. (1998), as abordagens com base em regras buscam descobrir o que é ‘teoricamente’ possível na língua, analisando unidades estruturais e classes da língua, e descrever como as palavras se combinam para formar frases e textos. Uma outra abordagem, chamada pelos autores de abordagem com base no uso da língua, permite investigar a ocorrência de estrutura(s) específica(s), previamente selecionada(s) pelo pesquisador, em diferentes contextos, podendo concentrar-se em um único texto ou conjunto de textos, um corpus.

Diariamente as pessoas têm contato com os mais variados tipos de texto que acontecem em diferentes níveis da língua: a leitura do jornal, uma conversa com um amigo, a escritura de um e-mail, por exemplo. A língua, assim como a matemática e outras ciências, é regida por regras que determinam quais combinações são possíveis entre seus símbolos para, no caso da língua, formarem frase providas de significado. Salvo algumas exceções como os poemas, que fazem parte da língua mas algumas vezes violam suas regras. Lingüistas como Chomsky discutem que o ser humano tem a capacidade de fazer uso infinito desse meio finito de ‘símbolos’, respeitando os princípios básicos das suas línguas naturais, o aspecto criativo da língua.

A análise do uso apresenta dois objetivos centrais, que seriam: avaliar o grau em que um padrão é encontrado e analisar os fatores contextuais que influenciam a variação de uso (Biber, Conrad et al., 1998). Em termos práticos, segundo o mesmo autor, pode-se

comparar a preferência do uso da estrutura *that-complement clauses* por um grupo de falantes **X** e da *to-clauses* por um grupo de falantes **Y** da língua inglesa.

A pesquisa com base em corpus, resumidamente, caracteriza-se por (Biber, Conrad et al., 1998: 4):

- ser empírica, analisando os padrões atuais de uso em textos naturais;
- utilizar *corpus* como base para análise;
- fazer uso extensivo computadores para análise, utilizando tanto técnicas automatizadas quanto interativas;
- depender tanto de métodos quantitativos quanto de qualitativos.

Às características da lingüística de corpus alia-se o uso do computador, que trouxe novos rumos para a manipulação do *corpus*, permitindo a execução de tarefas elementares, em termos de processamento de dados, como armazenamento e análise de um grande volume de dados, recursos de busca, recuperação, seleção e cálculo dos conteúdos de corpora com rapidez, se comparado ao trabalho manual. A Lingüística de Corpus vem sendo amplamente utilizada em Processamento de Linguagens Naturais (PLN) e também para o estudo de fenômenos lingüísticos da fonética, morfologia, sintática e semântica.

O uso do corpus para o estudo das palavras, a lexicologia, possibilitou a análise de associações lingüísticas e não-lingüísticas de uma palavra, os usos mais freqüentes de uma determinada palavra, o contexto em que ela é mais utilizada, ao invés de apenas

trazer o seu significado. Para a sociolingüística, Biber, Conrad et al. (1998) destacam o uso do corpus para estudos como a investigação de dialetos, padrões de co-ocorrência em diferentes registros. A lingüística diacrônica, que "estuda a língua em função de suas variações de lugar para lugar, de falante para falante e de um tempo para outro" (Saussure, 1977), pelo uso do 'corpus histórico', pode comparar o uso da língua ao longo dos tempos, bem como textos de diferentes autores em diferentes épocas. O uso de corpus, então, ampliou o horizonte da pesquisa lingüística.

2.1.5. A ligação entre a Lingüística Computacional e a Lingüística de Corpus

O aparecimento oficial do termo lingüística de corpus deu-se no livro *Corpus Linguistics*, (Aarts e Meijs, 1984). O uso de um computador para desenvolver análises com base em corpus implicou uma relação no desenvolvimento da lingüística de corpus e da lingüística computacional. "Lingüística computacional é o termo utilizado genericamente para denominar a investigação de linguagens humanas por meio de computadores, enquanto, na lingüística de corpus, esta investigação faz uso de um corpus e, atualmente, também de um computador (...)" (Rocha, 2000). Tanto a lingüística computacional quanto a de corpus oferecem bases metodológicas para a pesquisa lingüística, podendo ser aplicadas a diversos ramos de pesquisa como a fonética, sintaxe, semântica, sociolingüística, entre outras.

Em Leech (1992), a caracterização de ambas as áreas como metodologias de análise é apontada:

The only other branch of linguistics which, like corpus linguistics, refers to a tool or methodology rather than a subject-matter is computacional linguistics, defined as the investigation of language by means of computers.

O termo lingüística computacional é geralmente confundido com PLN (Processamento de Linguagens Naturais). O último, todavia, está mais relacionado com a inteligência artificial. Alguns dos problemas encontrados em PLN são solucionados pela análise e manipulação de corpora, e as soluções encontradas podem ser úteis para o estudo de fenômenos lingüísticos. (Para mais detalhes, ver Rocha, 2000)

2.1.6 Metodologia de corpus e o uso de corpora para a lexicologia

A lexicologia estuda o uso e o significado das palavras. O Dicionário Houaiss define lexicologia como "parte da lingüística que estuda o vocábulo quanto ao seu significado, constituição mórfica e variações flexionais, sua classificação formal ou semântica em relação a outros vocábulos da mesma língua, ou comparados com os de outra língua, em perspectiva sincrônica ou diacrônica.". A riqueza de exemplos autênticos que o uso de corpora traz para a pesquisa lexicográfica permite que se tenha uma noção mais acurada da freqüência, do uso e até do valor social de uma palavra ou do seu sentido.

A pesquisa lexicográfica, no que envolve a compilação da grande maioria dos dicionários, entretanto, vinha analisando apenas o uso padrão das palavras, formas e usos que já se tornaram tradicionais na língua, consagrados pela gramática tradicional, sem levar em conta situações reais de uso: como, quando, com que freqüência e em que contextos elas são utilizadas na língua.

"O dicionário descreve, primeiramente, as características formais das palavras: como são soletradas, pronunciadas, suas flexões e como se formam palavras derivadas e compostas. Num segundo momento, suas características semânticas – o que significam – e, finalmente, suas características combinatórias – como combinam com outras palavras." (Svensén, 1993, tradução da pesquisadora)

Segundo Othon Garcia (1993), em praticamente nenhuma língua conhecida uma palavra (significante) tem um só sentido (significado), salvo algumas exceções; as palavras são polissêmicas e, dependendo do contexto em que estão inseridas, mudam de valor. Pode acontecer de nem mesmo um dicionário dar conta de definir um significante de acordo com a intenção do autor, no caso da língua escrita, ou do falante, no caso da língua falada. Autores como Biderman (1981) afirmam que o léxico inclui conceitos lingüísticos e não-lingüísticos da nomenclatura de referentes do mundo físico e do cultural, estando por isso situado entre o *lingüístico* e o *extralingüístico*.

Com base nos dados expostos acima, pode-se afirmar que o uso de um corpus para a pesquisa lexicográfica pode mostrar as muitas discrepâncias entre os significados de uso de uma palavra e os tradicionais, que muitas vezes não são frutos de uma distorção regional, permitindo ao lexicógrafo apurar a freqüência e como (com que sentido), por exemplo, a palavra é empregada pelos falantes de uma língua, tendo o seu trabalho fundamentado em critérios léxico-estatísticos. Biderman (2001: 135, 136) apresenta uma lista de regionalismos dividida pelas regiões do Brasil:

"Amazônia: boto, curimatã, gaiola. Igarapé, progonga, pororoca, quatipuru, seringa tucunaré;

Sul do Brasil: arregalar, enriconar, bagual, bombachas, bomba de chimarrão, china, chiripá, guaiaca, mate, poncho, querência;

Nordeste do Brasil: aipim, caatinga, cajueiro, jangada, macaxeira, maruim, marruá, oitizeiro, saveiro;

Pantanal: anhuma, curicaca, seriema, tuiuiú, acuri, aguaçu, canjiqueira, peúva."

Ainda segundo a autora, para que um corpus seja utilizado na compilação de um dicionário, deve ter aproximadamente 10 milhões de palavras, com todas as modalidades de uso (texto e discurso, oral e escrito). Selecionam-se, então, as palavras que ocorreram 5 vezes para, em seguida, verificarem-se as que ocorreram com freqüência entre 1 e 5;

aquelas que apresentam freqüência 1 não são utilizadas.[2] A pesquisa com base em corpus possibilita a avaliação dos diferentes usos das palavras, em diferentes contextos, tarefa de que os dicionários cuja compilação não conta com o uso de um corpus, não conseguem dar conta por não contarem com a ajuda de um corpus de referência. Pode-se relacionar a incomplitude dos dicionários ao crescimento geométrico do léxico, fato acelerado devido também à globalização. Há de se considerar, entretanto, que não deve ser uma tarefa fácil selecionar uma média de 50.000 palavras de um acervo lexical; ela pode ser amenizada com o uso de um corpus, como já foi discutido anteriormente.

Sabe-se que um homem culto domina aproximadamente 25.000 palavras, incluindo o léxico ativo e o passivo. Se um dicionário é destinado ao grande público, é suficiente que tenha uma macroestrutura de 50.000 entradas. (Para mais detalhes, ver Biderman, 1981). O *Collins Cobuild Dictionary of English Language* (1987), por exemplo, de autoria de John Sinclair e de sua equipe (University of Birmingham), foi o primeiro dicionário desenvolvido com base em um corpus computadorizado, o *Bank of English*, corpus de referência com 450 milhões de palavras do inglês britânico e um subcorpus do inglês americano. O dicionário é destinado tanto a falantes nativos da língua inglesa quanto aos que falam o inglês como segunda língua. Em 1995 foi lançado o *Cambridge International Dictionary of English* (Cambridge University Press), também elaborado com base em corpus.

Hoje em dia, o trabalho com corpus é relativamente simples, em termos de tecnologia, pois já se conta com a ajuda de programas de manipulação de corpus, como o

[2] Os números aqui apontados são os ideais para a compilação de um dicionário comercial de 50 000 palavras.

utilizado na realização da presente pesquisa, o *WordSmith* (Scott, 1996). A maioria dos programas de manipulação de corpus apresentam uma lista de freqüência de todas as palavras presentes no corpus, a qual mostra o número de vezes que cada palavra aparece no corpus. No programa *WordSmith* as listas de freqüência podem ser geradas em ordem alfabética, ordem de ocorrência ou de freqüência. Tal programa tem três ferramentas básicas: *Wordlist* (lista de palavras), *Concord* (concord) e *Keyword* (palavras-chave). Tendo em vista os objetivos propostos para esta pesquisa, fez-se uso do *Concord*, programa concordanceador. Essa ferramenta permite ao pesquisador analisar como as palavras aparecem e se 'comportam' no corpus, bem como visualizar qualquer palavra do corpus no contexto em que ela ocorre, e suas combinações.

Dentre os programas de manipulação de corpus, além do *WordSmith* podem ser citados o TACT (Text- Analysis Computing Tools), o Corpus Workbench e o ICEUP (International Corpus of English Utility Program). A combinação do uso desses programas com o do corpus computadorizado permite ao pesquisador analisar um grande número de ocorrências de palavras a partir de textos que as pessoas lêem e escrevem diariamente.

CAPÍTULO 3 – METODOLOGIA

3.1 CONTEXTUALIZANDO A PESQUISA

3.1.1 Descrevendo o cenário da pesquisa

Quando se pensa em pesquisa com base em corpus, é preciso ter em mente algumas questões importantes no âmbito de pesquisa da Lingüística de Corpus. Uma delas é a representatividade do corpus. Sardinha (2000) afirma que “um corpus, seja de que tipo for é tido como representativo da linguagem, de um idioma, ou de uma variedade dele”. A representatividade encontra-se relacionada também ao tamanho do corpus; quanto maior, mais representativo, porque permite uma amostragem segura do que está sendo pesquisado. Algumas palavras têm a freqüência de ocorrência baixa; sendo assim, num corpus pequeno, a probabilidade de elas aparecerem é praticamente nula. Outras palavras apresentam sentidos diferentes, o que pode ser melhor analisado em um corpus de grande porte: sentidos mais e menos freqüentes. Pode acontecer de enunciações raras serem incluídas, enquanto outras bastante comuns ficarem de fora, a menos que se esteja trabalhando com uma língua morta, cuja quantidade de textos é fixa.

Rocha (2000) afirma que “as capacidades de armazenamento dos computadores da atualidade, aliadas a periféricos com capacidade de digitalização por meio de leitura ótica e aos dados que podem ser diariamente obtidos através do WWW, permitem coletar corpora muito grandes, contendo amostras de uma ampla gama de gêneros.”. Continua: “a quantidade de gêneros evita que formas relativamente raras assumam importância excessiva na amostra coletada.” O autor conclui dizendo que a inclusão de uma grande variedade de gêneros em um corpus permite a anulação das diferenças específicas aos gêneros textuais. O corpus deve conter uma variedade suficiente de textos, os quais devem adequar-se aos propósitos da pesquisa.

O ‘quanto maior melhor’ não pode ser definido em números específicos, uma vez que “o corpus é uma amostra de uma população cuja dimensão não se conhece (a linguagem como um todo).” (Sinclair, 1991, in Sardinha, 2000) Como não se podem estabelecer números específicos do tamanho ideal de uma amostra, a solução é tornar essa amostra o maior possível, o que a aproxima da população da qual foi extraída.

Verifica-se, assim, que a seleção de um corpus envolve pré-requisitos, de acordo com Sardinha (2000):

> *(1) Primeiramente, o corpus deve ser composto de textos autênticos, em linguagem natural. Assim, os textos não podem ter sido produzidos com o propósito de serem alvo de pesquisa lingüística. E não podem ter sido criados em linguagem artificial, tais como*

linguagem de programação de computadores ou anotação matemática.

(2) Em segundo lugar, quando se fala em autenticidade dos textos, subentende-se textos escritos por falantes nativos. Tanto assim que, quando este não é o caso, deve-se qualificá-lo, falando-se em corpora 'de aprendizes' ('learner corpora').

(3) O terceiro pré-requisito é que o conteúdo do corpus seja escolhido criteriosamente. Os princípios da escolha dos textos devem seguir, acima de tudo, as condições de naturalidade e autenticidade. Mas devem também obedecer a um conjunto de regras estabelecidas pelos seus criadores de modo que o corpus coletado corresponda às características que se deseja dele. Ou seja, o conteúdo do corpus deve ser selecionado a fim de garantir que o corpus tenha uma certa característica. (..)

(4) O último, a representatividade.

Tendo analisado o corpus escolhido para esta pesquisa com base nos critérios citados acima, verificou-se que ele estava de acordo com os pré-requisitos apontados, podendo ser utilizado para que se alcancem os objetivos propostos para esta análise. O corpus a ser utilizado nesta pesquisa é o do NILC (Núcleo Institucional de Lingüística

Computacional), dada a sua variedade de textos, suficientemente boa. Os textos estão divididos em três categorias: os corrigidos, os publicados para um grande número de leitores, têm um total de 32.590.000 palavras em 4.300 textos.

Os gêneros textuais são diversos: livros (de literatura brasileira; didáticos – biologia, química, física, história, geografia; enciclopédias; temáticos - arte, ciências, etc.); revistas; constituição brasileira e textos jurídicos; jornais. Ao contrário dos textos corrigidos, os semicorrigidos foram publicados para um número pequeno de leitores, ou nem mesmo foram publicados. São eles: contratos, relatórios, dissertações acadêmicas, etc. Os textos extraídos de jornais contabilizam a grande maioria da população do corpus; são eles dos cadernos do Jornal do Brasil e da Folha de São Paulo. A diversidade dos cadernos desses jornais possibilitou a análise de ocorrências do verbo *ficar* em diferentes contextos. Os textos não corrigidos totalizam 738.000 palavras, em 2.400 textos autênticos, que incluem redações, monografias, textos de publicidade, entre outros, produzidos por alunos do ensino médio e por universitários. Por fim, os textos semicorrigidos, com um total de 1.115.000 palavras em 238 textos. A partir da descrição do conteúdo apresentado pelo corpus do NILC, pode-se concluir que seus textos são autênticos; foram produzidos para outros fins que não a pesquisa lingüística, e escritos por falantes nativos, em linguagem natural.

O *ficar*, objeto de análise deste trabalho, foi selecionado por se tratar de um verbo polissêmico da língua portuguesa, e por apresentar um número significativo de ocorrências no corpus selecionado: 43525, em 42 formas diferentes (tabela 2). De todas

as suas ocorrências, relacionadas na tabela 2, será aqui analisada uma amostra com 500 delas, selecionadas aleatoriamente com a ajuda do programa de computador *WordSmith*, nos textos que envolvem um total de aproximadamente 33.328.000 palavras. O mesmo programa foi utilizado como ferramenta para o levantamento de dados, o que possibilitou também esse tipo de análise quantitativa com precisão.

Tabela 2 *Freqüência e proporção dos usos de ficar no corpus do NILC*

FORMA DO VERBO	NÚMERO DE OCORRÊNCIAS	% DO TOTAL DE FORMAS DE *FICAR* NO CORPUS
FICAIS	1	0,002297
FICARIAS	1	0,002297
FICASSES	1	0,002297
FICAVAS	1	0,002297
FICAREIS	3	0,006892
FICARES	3	0,006892
FICASTE	4	0,009189
FICAI	5	0,011487
FICÁSSEMOS	5	0,011487
FICARÁS	6	0,013784
FICAS	8	0,018379
FICARÍAMOS	16	0,036757
FICÁVAMOS	20	0,045946

FIQUEMOS	28	0,064325
FICAREI	30	0,06892
FICAREMOS	43	0,098785
FICARMOS	52	0,119461
FICASSEM	87	0,199867
FICASSE	87	0,199867
FICARA	116	0,266489
FICARIAM	234	0,537573
FICAREM	235	0,53987
FICAVAM	252	0,578924
FICADO	540	1,240552
FICARÃO	582	1,33704
FICO	663	1,523123
FICAVA	752	1,727584
FICARIA	795	1,826369
FIQUE	919	2,111236
FICANDO	1055	2,423672
FIQUEI	1172	2,692458
FICARÁ	1416	3,253004
FICAM	2781	6,388844
FICARAM	3318	7,622505
FICAR	7519	17,27354
FICOU	9790	22,49075

FICA	9970	22,90427
TOTAL DAS OCORRÊNCIAS	43525	100%

A tabela 2 mostra os diferentes usos do *ficar* no corpus de pesquisa. Tal levantamento estatístico objetivou analisar a freqüência do verbo no corpus do NILC, o que mostrou a relevância de se estudar o uso do verbo pela sua freqüência representativa. O total de ocorrências do verbo pode ser contrastado com o total das ocorrências na amostra de estudo, a fim de comparar esses números (ver tabelas 2 e 3). Comparando-se as tabelas 2 e 3 vê-se que as formas mantiveram a proporção de ocorrência e as formas com freqüência de número 1 não entraram na seleção. A forma *ficasse* apresenta freqüência de número 1 na amostra de estudo, enquanto que no total do corpus apresenta freqüência 87, assim como a forma *ficasse*, que também apresentou uma freqüência baixa, caindo de 87 para 2, e *ficara*, que baixou de 116 para 2.

Para explicar o porquê da exclusão das formas com freqüência de número 1 da seleção da amostra, pode-se citar Sardinha (2000), que discute que “em qualquer corpus, as formas de freqüência 1 (também conhecidas como ‘hapax legonema’) são a maioria. Baseando-se nesse fato, é possível afirmar que o léxico de freqüência baixa é o mais comum, isto é, a maioria das palavras de uma língua é composta de palavras que ocorrem poucas vezes. Em outras palavras, palavras de baixa freqüência tem uma probabilidade baixa de ocorrência (1 em 1 milhão, por exemplo), e já que elas formam a maior parte do vocabulário de uma língua, torna-se necessário amostras grandes para que elas possam

ocorrer.” A passagem explica o porquê da ausência das assim chamadas ‘formas raras do verbo’, mas a polissemia do verbo não está relacionada a sua forma (passado, presente, futuro), mas ao que o acompanha, adjetivos, advérbios e preposições, como se verificará na análise do corpus de estudo.

Tabela 3 *Freqüência e proporção dos usos de ficar na amostra*

FORMA DO VERBO	NÚMERO DE OCORRÊNCIAS	% NA AMOSTRA
FICASSE	1	0,2
FICARA	2	0,4
FICASSEM	2	0,4
FICO	3	0,6
FICARIAM	4	0,8
FICAVA	4	0,8
FICADO	7	1,4
FICAMOS	8	1,6
FICAREM	8	1,6
FICARÃO	11	2,2
FICANDO	12	2,4
FICARIA	13	2,6
FICARÁ	22	4,4
FICAM	28	5,6
FICARAM	56	11,20

FICAR	83	16,60
FICA	92	18,40
FICOU	137	27,40
TOTAL	500	100%

A busca por uma palavra específica, no contexto desta pesquisa o verbo *ficar*, foi feita utilizando-se o programa *Concord*, dentro do *WordSmith*. Como a quantidade desejada de ocorrências era 500, para que a seleção não fosse realizada manualmente utilizou-se o recurso *Horizons* do *Concord*, que permite fazer a seleção aleatória. Primeiramente, foi feita uma amostra com 16.000 ocorrências, que é o limite de busca do programa, dentro das 43.525 encontradas. Em seguida, uma ocorrência em cada 32, totalizando, assim, 500. Programas como o *WordSmith* também conhecidos como 'concordanceadores', permitem ao pesquisador buscar uma palavra específica em um corpus, e produzem listas de freqüência, listas por ordem alfabética, colocações lexicais e ocorrências da palavra escolhida, conforme exposto anteriormente. Pelo recurso da lista de palavras é possível saber a freqüência de todas as palavras do texto, dados estatísticos precisos e o número total de ocorrências de uma determinada palavra no corpus.

A análise dos sujeitos das ocorrências da amostra mostrou que, em 334 ocorrências, o verbo tinha como sujeito seres inanimados. Tal fato talvez esteja relacionado ao gênero dos textos que compõem o corpus selecionado para o desenvolvimento da pesquisa, que são em sua maioria jornalísticos.

(1) “Tiveram queda 29 setores e 8 *ficaram* estáveis.”

(2) “O América ficou plantado atrás, deixando claro desde cedo que um empate já seria considerado um bom resultado.”

Em oposição, foram encontradas 166 ocorrências que apresentaram o verbo acompanhado de um sujeito ‘ser animado’:

(3) “A fuga dos 98 presos do distrito foi evitada depois que o detento Carlos Henrique da Silva, 24 *ficou* entalado no meio do túnel.”

(4) “N. ficou ferido e uma passageira do Fiat, não identificada, morreu.”

Os dados apresentados vêm ao encontro da classificação do verbo *ficar* apresentada pela gramática gerativa. Tal gramática classifica esse verbo como sendo inacusativo, porque ele não c-seleciona o seu argumento externo (sujeito), não lhe impõe restrições selecionais, aceitando como argumento tanto seres animados (humanos) como não-animados. Dessa forma, o verbo permite múltiplas combinações lexicais entre o argumento interno e o argumento externo (sujeito e complemento).

Tendo-se analisado o verbo quantitativamente, levantando-se a característica dos tipos de sujeito que ele seleciona, o verbo será analisado no capítulo a seguir, dentro do contexto em que foi empregado, verificando-se se esse influencia no sentido daquele (cf.

Othon Garcia, 1993), o que levará à questão da polissemia da língua. Para a análise optou-se por considerar para a análise lingüística o contexto em que as expressões lingüísticas ocorrem, como e quando elas evocam um determinado significado, vindo ao encontro dos objetivos delimitados para a pesquisa; dentre eles, o de analisar o verbo *ficar,* com base no uso da língua, em diferentes contextos. Com isso, a pesquisadora poderá contrastar o uso do verbo no corpus escolhido, com as definições apresentadas pelo Dicionário Houaiss. A partir dos dados levantados na comparação entre o uso do verbo e as definições, buscar-se-á responder às questões levantadas, bem como verificar se o dicionário selecionado para esta análise dá conta do uso do verbo. Para tanto, as ocorrências do verbo *ficar*, encontradas na amostra selecionada para esta pesquisa, foram analisadas com a freqüência, colocações e todo o material necessário para caracterizar os padrões de co-ocorrência.

Nas seções a seguir serão apresentados os objetivos e hipóteses de trabalho, uma revisão bibliográfica sobre a polissemia e brevemente expor alguns conceitos da lingüística cognitiva apresentados por Harris (1992), a fim de se contextualizar o verbo *ficar* no cenário da pesquisa.

3.1.2 Objetivos da pesquisa e hipóteses de trabalho

- Objetivos Gerais

- Analisar o verbo com base no uso da língua dentro de diferentes contextos;

- Fazer um estudo dos sentidos assumidos pelo verbo selecionado;

- Objetivos Específicos

- Verificar se o dicionário selecionado para a pesquisa dá conta do uso tal qual verificado neste estudo, pois os critérios utilizados na elaboração dos dicionários nem sempre levam em conta o uso das palavras.

- Hipóteses de trabalho

As seguintes hipóteses direcionaram a realização dos objetivos da pesquisa:

1) A variável lexical[3] influencia o sentido do verbo (ou não).

2) É possível estabelecer um padrão entre as variações lexicais e as variações semânticas (ou não).

3.1.3 Definindo a polissemia

A polissemia é um fenômeno lingüístico presente em línguas naturais. Uma palavra é polissêmica quando comporta mais de uma significação, porém com um mesmo

[3] Entenda-se por variável lexical os padrões de co-ocorrência do verbo.

sentido de base, um lexema com mais de um semema. Nesse caso, o próprio contexto é que se encarrega de focalizar uma de suas interpretações.

Segundo Bréal (1897, Passim), citado em Rehfeldt (1980), “o acúmulo de significações de um vocábulo representa diversidade de aspectos da atividade intelectual e social”. E complementa: “compete à semântica o exame das causas que levam as palavras, uma vez criadas e providas de certo sentido, a restringir o significado, a estendê-lo, a transportá-lo de uma ordem de idéias a outra”.

Perini, em sua Gramática Descritiva do Português (1995), afirma:

> *A polissemia é uma propriedade fundamental das línguas humanas, que sem ela não poderiam funcionar eficientemente. Seria impraticável dar um nome separado a cada "coisa", incluindo aquelas que nunca vimos. Ao nos depararmos com um objeto nunca visto antes - digamos, um novo modelo de bicicleta - ficaríamos sem recursos para denominá-lo. Mas não é assim que a linguagem e a mente trabalham. Ao encontrar um objeto novo, tentamos imediatamente "reconhecê-lo", encaixando-o em alguma categoria já existente na memória (e na língua).*

Pustejovsky não coloca a existência ou a origem da polissemia em prova. Sua teoria, a do Léxico Gerativo, doravante LG, quer “apenas” acrescentar a idéia de que o

próprio léxico traz muitas informações sobre as interpretações e sentidos que uma palavra pode assumir.

A polissemia costuma se confundir com a homonímia, pois duas ou mais palavras são homônimas por carregarem a mesma forma escrita $^{e}/_{ou}$ fonológica. Pustejovsky considera a homonímia como uma ambigüidade contrastiva, já que há contraste entre as palavras, e traz um exemplo clássico em seu livro, o qual apresenta a palavra 'banco' em dois contextos distintos[4]:

(1) João está sentado no banco.

(2) João é cliente deste banco há anos.

(3) João trabalha no banco.

(4) O banco está precisando ser pintado.

(5) O banco ficou de me ligar ainda hoje.

Com muita clareza pode-se perceber que 'banco' (1) (artefato) e 'banco' (2) (instituição financeira) são palavras homônimas, pois seus significados não se relacionam diretamente. Temos banco-1 = artefato, feito para sentar e banco-2 = instituição financeira.

Nos exemplos de (1) a (5), Pustejovsky está-se referindo a um sentido específico que a palavra 'banco', em seu significado de instituição financeira, pode representar. Em

[4] Os exemplos foram adaptados a partir daqueles apresentados por Pustejovsky.

(3) há referência à instituição propriamente dita; em (4), ao prédio onde se localiza o banco e, em (5), às pessoas que nele trabalham.

Lobato (1977), discute que em alguns casos a polissemia pode estar mais relacionada à ambigüidade. Esse tópico deve ser discutido juntamente com a polissemia e a homonímia, uma vez que existem dois tipos de ambigüidade: a lexical (as próprias polissemia e homonímia) e a ambigüidade sintática, em que pelo menos duas estruturas gramaticais podem ser atribuídas à sentença. Na sentença abaixo, temos um terceiro tipo de ambigüidade, em que a lexical e a sintática estão presentes:

(6) Canto da sala.

(Podendo referir-se à primeira pessoa do singular do verbo cantar no tempo presente e ao ângulo formado entre duas paredes, na sala.)

3.1.4. A lingüística cognitiva

Contrapondo-se à teoria de Chomsky, o qual afirma que a linguagem é inata ao ser humano, a lingüística cognitiva prega que a linguagem é o produto de habilidades cognitivas, o reflexo do conhecimento enciclopédico/ lingüístico do falante, um 'sistema' convencionalizado para codificar intenções comunicativas. A análise de expressões lingüísticas deve considerar o contexto em que elas ocorrem, como e quando, como elas

evocam um determinado sentido, e podem ser utilizadas satisfatoriamente pelo falante de uma determinada língua. Os sentidos de frases e expressões são esquemas culturalmente divididos e são descritos com mais ou menos abstração, evocam diferentes significados, dependendo da intenção do emissor e da percepção do receptor. A capacidade dos falantes de uma determinada língua para construir e esquematizar as formas das expressões e seus significados é um fator central da competência lingüística.

A gramática de uma língua é considerada a junção das suas expressões e seus significados, em que 'significados' engloba as conceitualizações evocadas, incluindo funções comunicativas e aspectos extralingüísticos. A lingüística cognitiva enfatiza a busca por princípios que facilitem a explicação da variedade de expressões de uma determinada língua, uma vez que o significado delas está estreitamente relacionado com a compreensão do receptor, que também engloba o seu conhecimento de mundo.

Harris (1992) afirma que "tentativas tradicionais para entender o sentido de expressões de línguas naturais têm dividido o problema em duas partes: (1) como os significados podem ser caracterizados e (2) como os significados das palavras combinam para produzir o significado das sentenças (...) O significado das sentenças é então entendido como a união entre o sentido das palavras que as sentenças contém", sendo assim chamado de composicional porque o sentido das sentenças depende da combinação do sentido das palavras. As palavras formam um conjunto que vai definir o sentido das sentenças.

CAPÍTULO 4 – RESULTADOS: DESCRIÇÃO E ANÁLISE

4.1 A coleta de dados

De acordo com o exposto no Capítulo 3, os dados foram selecionados utilizando-se as ferramentas de um programa de manipulação de corpus, o *WordSmith*; a principal ferramenta empregada foi o *Concord*, que permitiu a busca específica pelo verbo *ficar* no corpus do NILC.

Sabe-se que para a compilação de um dicionário comercial, de 50 000 palavras, a freqüência de ocorrência de uma palavra, colocação, dever ser de no mínimo 5, num corpus com aproximadamente 10 milhões de palavras (Biderman, 2001), de acordo com os dados apresentados no Capítulo 2. Como os resultados da análise do verbo *ficar* seriam comparados com os sentidos apresentados pelo Dicionário Houaiss da língua portuguesa (2001), optou-se por trabalhar com o corpus organizado pelo NILC, dada a sua variedade de gêneros textuais e sua representatividade.

4.2 Exploração do corpus

A abordagem com base em corpus serviu de guia para a análise do verbo, já que tal área da lingüística prega que o contexto é peça importante para a análise de expressões

lingüísticas, e os sentidos de frases ou expressões podem evocar diferentes significados, dependendo das intenções do emissor e do receptor. (Para mais detalhes, ver 3.1.4)

O programa responsável pela busca do verbo escolhido, *Concord* (*concordanceador),* processou e apresentou as ocorrências do verbo no corpus; no caso desta pesquisa, o verbo *ficar* foi o foco da análise lingüística aqui apresentada. No Capítulo 3 foi exposto que, a fim de se chegar a uma amostra com um número viável e suficiente de ocorrências, na primeira busca feita no corpus foi utilizando o recurso *WordList*, a partir do qual foi possível observar-se a freqüência do verbo pré-selecionado, o número total da sua ocorrência no corpus. Confirmada a freqüência, a etapa seguinte envolveu uma busca específica pelo verbo *ficar* em suas diversas formas.

A primeira análise da amostra de 500 ocorrências objetivou fazer um levantamento das colocações do verbo, suas variações lexicais, para caracterizar a sua polissemia: que palavra aparecia imediatamente à esquerda e que palavra aparecia imediatamente à direita do verbo, utilizando-se números e outros dados, de modo a caracterizar as acepções encontradas como colocações estáveis e suficientemente freqüentes no corpus para exigir uma menção explícita em um dicionário. Para a análise das acepções do verbo, suas variáveis semânticas, fez-se necessário observar o contexto em que ele aparecia em alguns casos. A análise das variáveis semânticas baseou-se nos sentidos apresentados no Dicionário Hoauiss.

Para efeitos de análise das variáveis lexicais do verbo, as colocações foram assim classificadas:

(1) SV- sintagma verbal

(2) SPrep- sintagma preposicional

(3) SAdj- sintagma adjetival

(4) SAdv- sintagma adverbial

(5) SN- sintagma nominal

Foram analisadas as ocorrências em que o verbo *ficar* aparecia, seguido das seguintes classes de palavras: preposição, adjetivo ou advérbio. As ocorrências em que o verbo aparecia seguido de um substantivo, fosse ele próprio ou comum, não foram excluídas da análise, e ficaram os substantivos classificados como SN. Na segunda parte da exploração foram analisadas as variáveis de sentido, dentro do contexto em que apareciam, para se chegar ao aspecto semântico do verbo e se poder comparar o sentido encontrado com aqueles apresentados pelo Dicionário Houaiss. Objetivou-se também analisar se um mesmo padrão de variável lexical da palavra sob investigação indicava um mesmo sentido ou sentidos diferentes.

Um dos resultados encontrados na análise do verbo foi em relação à classificação das variantes lexicais SAdj. Tais variantes representam 32% (tabela 5) do total das ocorrências da amostra. Dentre as entradas lexicais apresentadas pelo Dicionário Houaiss para o verbo *ficar*, há aquela em que *ficar* é descrito como sendo um *verbo-suporte* (*ficar*

gordo), o que seria, nesse caso, segundo a gramática tradicional, verbo de ligação (Cunha, 1970):

Os VERBOS DE LIGAÇÃO (ou copulativos) servem para estabelecer a união entre duas palavras ou expressões de caráter nominal. Não trazem propriamente idéia nova ao sujeito: funcionam apenas como elo entre este e o seu predicativo.

Observem-se os exemplos a seguir:

(6) "O ajudante do motorista, Edson Damasceno, 71, *ficou* preso nas ferragens do caminhão (...)"

(7) "A São Paulo que o adolescente conheceu nos anos 40, ao redor de Perdizes, *ficou* irreconhecível."

As variações lexicais ficar+SAdj foram então consideradas com o sentido de verbo de ligação. A perífrase do verbo *ficar*+particípio recebeu tratamento como se estivesse o verbo seguido de um adjetivo pois, segundo Travaglia (1994):

Na perífrase "ficar+particípio", o particípio vale por um adjetivo e o verbo ficar funciona como verbo relacional exatamente como em construções do tipo "ficar+adjetivo":

(8) “Uma das quatro pistas ascendentes da Imigrantes *ficou* fechada até às 9:15.”

(9) “Pista da marginal *ficou* interditada à noite.”

Nos exemplos acima, *ficar* faz a ligação entre o sujeito e o seu predicativo, confirmando a hipótese do verbo de ligação. Ainda assim, o verbo pode, em alguns casos, assumir o sentido do verbete número 4 do Dicionário Houaiss: “continuar algum tempo em determinada atitude, gesto, posição, situação ou estado; manter-se”. Nas ocorrências ‘ficar interditada’, ‘ficar fechada’, ‘ficar internado’, ‘ficar preso’, etc., a expressão tem um marco durativo, dando ao receptor a idéia de que a ação ‘subsiste’ por um tempo. O mesmo pode ser dito da expressão ‘ficar ferida’ (10), pois subentende-se que o sujeito da frase encontrava-se ferido e assim continuou, mas não é o caso das ocorrências da amostra pois os sujeitos das frases *ficaram feridos* em decorrência de acidentes sofridos pelos mesmos, seria até mesmo a idéia de mudança de estado. Ficaria difícil classificar os adjetivos de acordo com o sentido que cada um evoca no verbo *ficar*, o que poderia ser, além de tudo, controverso, uma vez que os adjetivos não são classificados como os advérbios. Os adjetivos, de acordo com a gramática tradicional, fazem parte de uma classe de palavras sem subdivisões, o que é sem dúvida discutível, mas não entra no mérito da questão da pesquisa. Optou-se, então, por se manter a classificação apresentada tanto pela gramática tradicional quanto pelo Dicionário Houaiss.

(10) “Entre sexta-feira e sábado, 12 pessoas morreram e 25 *ficaram* feridas em acidentes de trânsito na Bahia.”

Tendo sido esclarecidos os tratamentos dispensados para as duas variáveis lexicais SN e SAdj, passemos para a apresentação dos dados estatísticos da amostra. As conclusões preliminares incluíram apenas a análise das variáveis lexicais das ocorrências, que se encontram resumidas na tabela 4, a seguir. As ocorrências em que o verbo aparecia seguido de um sintagma nominal aparecem na tabela por razões de levantamento estatístico; no momento da análise das variáveis semânticas, porém, foram excluídas por questões expostas anteriormente.

Tabela 4- *Variáveis lexicais do verbo ficar*

ELEMENTO À ESQUERDA	VERBO	ELEMENTO À DIREITA	NÚMERO DE OCORRÊNCIAS
SN	Ficar	Sadj	123
SN	Ficar	Sadv	53
SN	Ficar	Sprep	158
SN	Ficar	SV	18
SN	Ficar	SN	30
SV	Ficar	SAdj	25
SV	Ficar	SAdv	5
SV	Ficar	SPrep	19
SV	Ficar	SV	4
SV	Ficar	SN	3
SAdj	Ficar	SAdj	5

SAdj	Ficar	SAdv	1
SAdj	Ficar	SPrep	3
SAdj	Ficar	SN	2
SAdv	Ficar	SAdj	5
SAdv	Ficar	SAdv	12
Sadv	Ficar	SPrep	4
SAdv	Ficar	SV	2
SAdv	Ficar	SN	4
Sprep	Ficar	SAdj	2
Sprep	Ficar	SAdv	1
Sprep	Ficar	SPrep	7
SPrep	Ficar	SV	3
TOTAL DA AMOSTRA			500

Resumidamente, os dados analisados encontram-se assim divididos:

Tabela 5 *Números e percentuais da amostra*

	SPrep	SAdj	Sadv	SV	TOTAL
FICAR +	174 (34,8%)	160 (32%)	83 (16,4%)	39 (7,8%)	456

Os dados da análise da variação lexical do verbo *ficar* mostram que a presença do verbo no corpus é marcada, na maioria das vezes, pela combinação verbo+SPrep, seguida pela combinação verbo+SAdj, o que indica uma forte tendência do uso desse

verbo como sendo de ligação. O sentido do verbo depende dos elementos que se encontram à sua esquerda ou direita, é deslexicalizado, ou seja, não tem sentido sozinho. Quando seguido da combinação verbo+SPrep, seu sentido vai variar de acordo com a preposição. Um dos significados evocados, o de permanecer em algum lugar, pode ser evocado também pela combinação verbo+SAdv. Tais combinações evocam também outros significados, como veremos a seguir. Comparando-se os números, vê-se que são poucas as ocorrências nas quais o verbo combina-se com um outro verbo.

De todos os adjetivos com que o verbo combina, o maior número de ocorrências ficou por conta de *ficar+ferido*. Tal fato certamente está estritamente relacionado ao registro textual predominante do corpus, o jornalístico. Outros adjetivos com freqüência expressiva na amostra foram: cheio, vazio, exposto, paralisado, caótico e congestionado, congelado.

Os advérbios 'abaixo' e 'acima' foram os que apareceram com mais freqüência e estavam relacionados, na maioria da vezes, a preços, juros, cotações. Outras combinações freqüentes são *ficar+fora* e *ficar+dentro*, que evocaram no verbo o sentido de estar situado, como veremos a seguir, na análise semântica das variantes lexicais.

A expressão *ficar no ar*, com freqüência 2 na amostra, apareceu relacionada à exibição de filmes (*Agora, enquanto não acha que é o momento adequado (...) o que fica no ar é um slide com a foto de Fábio Jr.*). Já a expressão *ficar claro*, com 6 ocorrências na amostra, aparece relacionada a algum evento, (*A seleção está começando a acertar.*

Isso ficou claro durante o treino de ontem.). As duas referidas expressões, entretanto, não estão listadas no Dicionário Houaiss, como veremos mais adiante na análise semântica.

Um dos sentidos mais recorrentes na amostra para o verbo *ficar* foi o de permanecer, continuar a estar em algum lugar (Hoauiss 2001). Tal sentido é evocado, entre outros, pelos verbos que acompanham *ficar*, que geralmente aparecem no gerúndio, dando ao verbo tal sentido (*Creio que em vez de ficar brigando com o câmbio (...)*).

Depois da análise da variação lexical do verbo *ficar*, foi feita a análise semântica, para verificar se as variáveis lexicais mantinham alguma regularidade com as variáveis semânticas e se as mesmas combinações apresentavam o mesmo sentido. As variáveis semânticas foram analisadas a partir do sentido que o verbo evocava no leitor. Em seguida, o sentido encontrado foi comparado com o que é apresentado pelo Dicionário Houaiss. A partir dessa análise foi feita a tabela 6: os sentidos foram classificados de acordo com o número da entrada lexical no dicionário. Do contrário, para os sentidos não listados, foram criadas novas classificações. As acepções apresentadas pelo citado dicionário encontram-se indicadas pelo nome do dicionário seguido pelo número da entrada lexical, Houaiss 4, Houaiss 2, e assim por diante.

Tabela 6: *Sentidos de ficar na amostra*

Acepções	Número de ocorrências na amostra	EXEMPLOS NO CORPUS
Continuar algum tempo em determinada atitude, gesto, posição, situação ou estado, manter-se (Houaiss 4)	69	Os investidores japoneses *ficaram* cautelosos.
Estar situado, localizar-se (Houaiss 5)	55	A área, de 900 metros, *fica* na rua 25 de Janeiro.
Montar a ou atingir (determinada quantia); custar, importar (Houaiss 22)	42	A desvalorização cambial *ficou* em 41,05%.
Permanecer num lugar, continuar a estar num lugar (Houaiss 1)	41	Exército *fica fora* das praias à noite.
Tomar posse ou apoderar-se (Houaiss 19)	25	A corretora Schahin Cury *ficou* com todo o lote.
Ser tomado ou considerado como (Houaiss 13)	9	Outra novidade do Bios, (como ficou popularmente conhecido)...
Responsabilizar-se por, abonar (Houaiss 25)	7	A Igreja *ficou* sob a tutela dos reis da França.
Ficar claro	6	*Ficou claro* que é necessário abordar não somente as decorrências do problema.

Ser excluído, não ser contemplado (ficar de fora/ Houaiss 36)	4	*Ficaram de fora* dessa etapa da isonomia os servidores da tabela 1.
Restar, sobrar (Houaiss 10)	2	*Ficou* só o Alferes, "louco e pobre."
Ser adiado, transferido, ficar para (Houaiss 16)	2	A decisão *ficaria*, então, para depois da convenção.
Ficar no ar	2	Eles *ficam no ar* até o lançamento do carro.
Ficar por um fio	1	Foram necessárias seis horas de reunião para salvar o acordo que (...) *ficou* de novo *por um fio*.

A tabela 6 mostra que o maior número de ocorrências da amostra ficou por conta do elucidado no Houaiss pela entrada de número 4, seguido pelo Houaiss 5, 22 e 1. Se comparados com o uso do verbo como 'verbo de ligação', cada um representa aproximadamente ¼ do total dessas ocorrências. As expressões *ficar claro, ficar no ar* e *ficar por um fio* não estão listadas pelo Dicionário Houaiss. Outras acepções, que veremos adiante, com ocorrência zero na amostra, são classificadas pelo dicionário em questão.

O número de ocorrências em que *ficar* evocou o sentido de ‘custar’ (Houaiss 22) surpreendeu no sentido de ter superado o sentido de ‘permanecer’ (Houaiss 1), em apenas 1 ocorrência. Esse número pode estar relacionado aos gêneros textuais que compõem o corpus do NILC, que são na sua maioria jornalísticos. Mas a real justificativa desse número significativo de ocorrências encontra-se no fato de que boa parte dos textos da amostra foram retirados do caderno de economia da Folha de São Paulo. O sentido, em Houaiss 22, depende do uso de uma preposição:

(11) “Assim, o acumulado desde julho, quando o Real foi criado, ficará entre 21,84% e 22,44%.”

Ou do uso de um advérbio:

(12) “Segundo Zogbi, o consumo de papel na imprensa até dezembro ficará 22% acima do registrado em 93.”

Outra conclusão a que se chegou foi que somente a variante lexical SAdj mantém a mesma variação semântica, conforme exposto no início do Capítulo, e evoca o uso do verbo *ficar* como sendo de ligação. Em alguns casos, porém raros, somente o contexto pode resolver a questão da polissemia do verbo. Se considerarmos:

(13) “No serviço público de nosso país, onde, além do salário, o emprego também faz parte dos direitos adquiridos, qualquer redução de despesas *fica* inviabilizada.”

O leitor, ao se deparar com o exemplo (13), poderia chegar à conclusão de que a redução de despesas sempre foi inviável, e assim permaneceu. Se acrescentarmos, porém, a expressão temporal omitida da frase no primeiro momento, exemplo (14), o leitor percebe que a intenção do emissor era evocar o sentido de que somente em momentos de crise a redução de despesas fica inviabilizada. A simples omissão de parte da frase, tal qual está no corpus, fez com que o verbo mudasse de sentido: o contexto, nesse caso, influencia no sentido do verbo.

(14) “No serviço público de nosso país, onde, além do salário, o emprego também faz parte dos direitos adquiridos, qualquer redução de despesas em momentos de crise *fica* inviabilizada.”

As ocorrências da combinação ‘ficar + SAdv’, com um total de 74 (tabela 5), variaram de acordo com a natureza do advérbio. Os advérbios de intensidade, seguidos por um adjetivo, mantiveram o mesmo sentido da combinação ‘ficar + SAdj’, como mostra o par de frases (15) e (16):

(15) “As chuvas na manhã de ontem fizeram o trânsito (...) ficar mais caótico.”

(16) “Córregos transbordaram e o trânsito *ficou caótico*.”

Os advérbios de lugar, conseqüentemente, atribuíram ao verbo o sentido Houaiss 5 (do dicionário objeto de análise): “estar situado; localizar-se.”

(17) “Exército *fica fora* das praias à noite.”

Sendo assim, pode-se afirmar que o sentido do verbo atribuído pelo advérbio depende da natureza do advérbio, pelo fato de *ficar* ser um verbo deslexicalizado. Em alguns casos, o advérbio de lugar pode atribuir o sentido citado acima, e em outros, o sentido de ‘custar’. O mesmo acontece com as preposições *em* (*no* e *na*) e *entre* (19), (20), que evocam os mesmos sentidos daqueles dos advérbios de lugar. Já a preposição *com* (18), dá ao verbo o sentido Houaiss 19: ‘tomar posse, apoderar-se’, sendo assim considerada uma regularidade assim como ‘ficar+SAdj’.

(18) “E o PFL que dá 100 *fica com* três ministérios?”

(19) “Brizola *ficou em* seu apartamento, em Copacabana, onde torceu para o Brasil longe da presença da imprensa. “

(20) “O tabuleiro de Dinha *fica no largo* de Santana (...).”

Na amostra analisada foram encontradas algumas expressões do verbo *ficar*, comuns na língua portuguesa, porém não relacionadas pelo Dicionário Hoauiss. O dicionário apresenta usos como: *ficar ao pintar (???), ficar atrás de, ficar de bem, ficar de fora, ficar de mal, ficar mal com, ficar por isso mesmo, ficar sobrando*. Na amostra analisada, foram encontradas ocorrências das expressões *ficar atrás de* e *ficar de fora*. Outras expressões recorrentes na língua portuguesa, todavia, não aparecem listadas no dicionário. São elas: ficar no ar, 2 ocorrências na amostra, e ficar claro, 6 ocorrências na amostra.

(21) "Os outros três filmes destacam as opções de cores, o design e o preço baixo. Eles *ficam no ar* até o lançamento do carro."

(22) "O próprio argumento de que 'não é dando comida que se resolverá o problema da fome no Brasil' foi colocado em seu devido lugar. *Ficou claro* que é necessário abordar não somente as decorrências do problema, mas também, e principalmente, os fatores que o determinam."

A expressão *ficar no ar* (21) evoca o sentido de permanecer, continuar, acepção apresentada para o verbo, pelo dicionário sob análise. Seu significado, contudo, envolve a idéia da transmissão, seja via rádio ou televisão, junto à idéia de permanência e continuidade. Ao se analisar o exemplo (22), em que aparece a ocorrência de *ficar claro*, nota-se que tal expressão implica que algo foi entendido por uma ou mais partes, ou então que foi esclarecido.

Com apenas uma ocorrência na amostra, a expressão *ficar por um fio* dá ao contexto em que está inserida o sentido de risco iminente:

(23) "Foram necessárias seis horas de reunião para salvar o acordo que (...) *ficou de novo por um fio*."

Outra acepção do verbo evidenciada pelo corpus, foi aquela com o sentido de restar, verbete número 10 do Houaiss:

(24) "Na hora do desenlace, frieza geral. *Ficou* só o Alferes, 'louco e pobre' - que contava com o apoio do povo e da tropa, em vão. Seu martírio foi ato repleto de conseqüências"

O termo *só* funciona tanto como um adjetivo (diz-se do ou o que está, momentaneamente ou não, desacompanhado, separado de outro(s))[5], como um advérbio (apenas, somente, unicamente)[6]. Seja como adjetivo ou advérbio, o termo atribui ao verbo da frase (24) o sentido de restar, estar sozinho.

Nos contextos analisados nota-se que o sentido do verbo está diretamente relacionado ao complemento que vem à sua direita. A combinação 'ficar+SAdj' pode ser altamente polissêmica, se a função do verbo na frase não for considerada 'verbo de

[5] Houaiss, 2001:2588.
[6] Idem.

ligação'. No caso de 'ficar + SAdv', vê-se claramente o efeito do tipo do advérbio sobre o sentido adquirido pelo verbo, uma vez que "essas palavras se juntam a verbos, para exprimir circunstâncias em que se desenvolve o processo verbal (Cunha, 1970)". Na maioria dos casos, a polissemia do verbo se dá pelos seus 'acompanhantes', pelos complementos que estão à sua direita. A questão do contexto serve mais para desambigüisar as sentenças.

Quando acompanhado por um advérbio de lugar, atribui-se o verbete 5 do Houaiss, o de localizar-se; para o advérbio de modo, são atribuídos dois sentidos, 4 e 13: o de manter-se e ser considerado como. O advérbio de exclusão, *só,* atribui o sentido descrito pelo verbete 10, restar. O mesmo acontece com a combinação 'ficar+SPrep'. O sentido do verbo dependerá do que vem depois da preposição, e os sentidos assumidos são os apresentados pelos verbetes: 1, permanecer (num lugar), 4 e 5, acima referidos, e 22, atingir uma quantia, custar.

Observou-se, também, que os gêneros textuais mais freqüentes para a investigação do verbo *ficar* na amostra deram-se nos textos jornalísticos, por constituírem estes a maioria no corpus. O gênero textual, entretanto, não exerce influência sobre o sentido do verbo, a não ser quando se considera o caso do sentido 'custar'. De acordo com o exposto anteriormente, sua alta freqüência da amostra pode estar relacionada aos gêneros textuais mais freqüentes no corpus e ao fato de ter sido selecionado um grande número de ocorrências do caderno de economia da Folha de São Paulo.

A questão sobre a influência do gênero textual no sentido do verbo mereceria maior atenção e um estudo mais aprofundado para que se pudesse chegar a outras conclusões que não as apresentadas nesse estudo, bem como os sentidos mais comuns em cada gênero.

No próximo capítulo, na Conclusão, há uma análise a respeito dos objetivos e hipóteses de trabalho levantados para esta pesquisa: um comentário geral sobre os resultados encontrados na análise.

CONCLUSÃO

A Lingüística de Corpus, como foi discutido ao longo dos Capítulos, proporciona uma investigação abrangente, confiável, pois o corpus representa um recorte de língua – escrita ou falada –, o que possibilita a busca por palavras ou expressões específicas em praticamente qualquer língua, bastando para isso a existência de um corpus da língua que se deseja estudar. Os programas de manipulação de corpus permitem uma busca rápida pela palavra/ expressão objeto de estudo, bem como um levantamento estatístico seguro de listas de freqüência de palavras no corpus, por exemplo. No caso dessa pesquisa os objetivos delimitados envolviam o estudo do uso do verbo *ficar* em língua portuguesa.

Sabe-se que, hoje em dia, muitas são as pesquisas em lingüística que se baseiam no uso da língua para alcançar os seus objetivos; os resultados se aproximam da realidade dos falantes e possibilitam analisar diferentes usos que surgem freqüentemente na língua oral e escrita. Como esta pesquisa buscou analisar o uso do verbo *ficar* em língua portuguesa, optou-se por essa metodologia de pesquisa, a com base em corpus, a pesquisa com base no uso da língua.

Os resultados encontrados mostraram algo que de certa forma já era esperado: um dos usos mais freqüentes de ficar é como verbo de ligação. A análise mostrou, entretanto, que tal tratamento dispensado para o verbo não é o ideal, pois a afirmativa de que o verbo serve como elo entre o sujeito e o seu predicativo nem sempre reflete o sentido do verbo na frase. Em casos como *ficar ferido* o verbo traz uma idéia nova ao sujeito que pode ser de mudança de estado: alguém estava são e em seguida não estava mais, ou simplesmente de permanência: alguém estava ferido e assim continuou.

Um dos objetivos propostos era verificar se os diferentes sentidos assumidos pelo verbo encontravam-se relacionados com o gênero textual em que *ficar* aparecia. Já a partir da análise das variáveis lexicais percebeu-se que o sentido do verbo está estreitamente relacionado ao seu ‘acompanhante`. A confirmação da afirmativa de que o sentido está relacionado ao gênero textual dependeria de um estudo mais aprofundado para também fazer um levantamento de quais gêneros evocam quais sentidos no verbo o que conseqüentemente causaria (ou não) o maior uso de determinados `acompanhantes` para um determinado gênero textual. A partir desse estudo pode-se observar que o gênero textual influencia na escolha dos acompanhantes do verbo; termos relacionados à economia, ou ao cotidiano e política foram mais freqüentes, em conseqüência do grande número de textos jornalísticos presentes na amostra. Mas o estudo que se fez da relação gênero-`acompanhante`-sentido foi superficial.

A análise das variáveis semânticas mostrou o alto nível polissêmico do verbo e a existência de expressões não citadas pelo Dicionário Houaiss, o que evidencia a importância do uso de um corpus para o estudo do uso da língua; no caso de um dicionário, a lexicografia. Com o corpus foi possível comprovar a existência dos usos evidenciados pelo dicionário, ainda que alguns não tenham aparecido na amostra e outros, da amostra, não estejam listados na edição estudada. Por não levar em conta o uso das palavras em sua compilação, os dicionário comerciais pecam por exceder em algumas acepções das palavras, que às vezes não são mais usadas, e deixar de lado outras, bastante empregadas no uso da língua. De um modo geral, o dicionário estudado dá conta do uso do verbo, não citando apenas três expressões: *ficar no ar*, *ficar claro* e *ficar por um fio*. Sabe-se, entretanto, que nem todos os registros de uso de uma palavra podem ser listados por um dicionário. A seleção deveria ser feita seguindo os critérios de mais uso ou menos uso de um determinado registro para não acontecer de expressões mais recorrentes da língua serem citadas enquanto outras menos freqüentes são citadas.

A busca por um padrão de sentido entre a variável lexical e a variável semântica mostrou que não é possível estabelecer-se totalmente um padrão entre tais variáveis, pois uma mesma preposição pode evidenciar diferentes sentidos. Dependendo da classe de palavras, pode-se estabelecer um padrão, como é o caso dos adjetivos, que dão ao verbo *ficar* a conotação de verbo de ligação, e dos advérbios que, por serem classificados, evocam sentidos, de acordo com a sua classificação. A variação de sentido do verbo *ficar* é uma propriedade dos verbos delexicalizados que têm um

correlato lexical, sendo assim, *ficar claro* equivale a clarear, *ficar fora* a sair, *ficar inviabilizado* a inviabilizar-se e assim por diante.

BIBLIOGRAFIA

ABERCOMBRIE, D. (1965) *Studies in phonetics and linguistics.* Londres: Oxford University Press

BIBER, Douglas. CONRAD, Susan e REPPEN, Randi. Corpus Linguistics: Investigating Language Structure and Use. Cambridge: Cambridge University Press, 1998.

BIDERMAN, Maria Tereza. A estrutura mental do léxico in *Estudos de Filologia e Lingüística*. São Paulo: Edusp, 1981, p138.

_____. Os dicionários na contemporaneidade: arquitetura, métodos e técnicas in *As Ciências do Léxico: Lexicologia, Lexicografia, Terminologia*. Campo Grande: Editora UFMS, 2001, pp. 131-151.

CUNHA, Celso. Gramática do Português Contemporâneo. Belo Horizonte: Editora Bernardo Álvares, 1970.

FOLTRAN, Maria José & WACHOWICZ, Teresa. *The Generative Lexicon* – Cad.Est.Ling.,Campinas, (39): 151-162, Jul/Dez. 2000.

GARCIA, Othon M. Polissemia e Contexto, *in Comunicação em prosa moderna.*Rio de Janeiro: Editora da Fundação Getúlio Vargas, 1996. 13a ed.

HARRIS, C. Connectionism and cognitive linguistics. In: Noel Sharkey (org.). *Connectionist natural language processing*. Oxford, Intellect, 1992.

HOUAISS, Antônio, VILLAR, Mauro de Salles, FRANCO, Francisco Manoel de Mello (orgs). Dicionário Houaiss da Língua Portuguesa. Rio de Janeiro: Editora Objetiva, 2001. 1ª edição.

JOHNS, Tim. If our Descriptions of Language are to be Accurate... A footnote to Kettemann. EISU, University of Birmingham, 1998.

LEECH, Geoffrey. Corpora and theories of linguistic performance. In: Jan Svartvik (ed.). *Directions in Corpus Linguisitics*, pp. 379 – 397. Berlim, Mouton de Gruyter, 1992.

MANNING, C. e Schütze, H. Foundations of statistical natural language processing. Cambridge, Mass., The MIT Press, 2000.

MIOTO, Carlos. LOPES, Ruth e SILVA, Maria Cristina. Manual de Sintaxe. Florianópolis, Insular, 1997.

PERINI, Mário A. – Gramática descritiva do português – Editora Ática. 4ª ed. São Paulo. 2002.

PUSTEJOVSKY, James – The Generative Lexicon – Cambridge (MA): MIT Press. 1995.

REHFELDT, Gládis Knak – Polissemia e Campo Semântico – estudo aplicado aos verbos de movimento. EDURGS/FAPA/FAPCCA. Porto Alegre – RS. 1980.

ROCHA, Marco. Métodos com base em corpus no processamento de linguagens naturais.Não publicado.

SARDINHA, Berber, A. P. (1999) Beginning portuguese corpus linguistics exploring a corpus to teach portuguese as a foreign language. In: *D.E.L.T.A.*, vol. 15, nº2, pp. 289-99. PUC/SP.

SARDINHA, Berber, T (2000) Lingüística de corpus: histórico e problemática. In: *D.E.L.T.A.*, vol.16, n 2, pp. 323-67. PUC/SP.

_____. O que é um corpus representativo? In: Direct Papers n 44. http://www.direct.f2s.com. Página acessada no dia 21/01/04 às 15:30h.

SINCLAIR, John M. The automatic analysis of corpora. In: Jan Svartvik (ed.). *Directions in Corpus Linguisitics*, pp. 379 – 397. Berlim, Mouton de Gruyter, 1992.

SMITH, George W. Computers and Human Language. Oxford, Oxford University Press, 1991.

SVENSÉN, Bo. Practical Lexicography, pp.1-9, 40-63. Oxford, Oxford University Press, 1993.

THOMAS, Jenny e SHORT, Mick (eds.). Using Corpora for Language Research. London, Longman, 1996.

TRAVAGLIA, Luiz Carlos. O Aspecto Verbal no Português: a Categoria e sua Expressão. Uberlândia: Edufu, 1994.